Simple Statistical Probabilistic Forecasts of the Winter NAO

RICHARD J. HALL

School of Geography and Lincoln Centre for Water and Planetary Health, University of Lincoln, Lincoln, United Kingdom

ADAM A. SCAIFE

Met Office, Hadley Centre, Exeter, United Kingdom

EDWARD HANNA

School of Geography and Lincoln Centre for Water and Planetary Health, University of Lincoln, Lincoln, United Kingdom

JULIE M. JONES

Department of Geography, University of Sheffield, Sheffield, United Kingdom

ROBERT ERDÉLYI

Solar Physics and Space Plasma Research Centre, School of Mathematics and Statistics, University of Sheffield, Sheffield, United Kingdom

(Manuscript received 5 July 2016, in final form 20 March 2017)

ABSTRACT

The variability of the North Atlantic Oscillation (NAO) is a key aspect of Northern Hemisphere atmospheric circulation and has a profound impact upon the weather of the surrounding landmasses. Recent success with dynamical forecasts predicting the winter NAO at lead times of a few months has the potential to deliver great socioeconomic impacts. Here, a linear regression model is found to provide skillful predictions of the winter NAO based on a limited number of statistical predictors. Identified predictors include El Niño, Arctic sea ice, Atlantic SSTs, and tropical rainfall. These statistical models can show significant skill when used to make out-of-sample forecasts, and the method is extended to produce probabilistic predictions of the winter NAO. The statistical hindcasts can achieve similar levels of skill to state-of-the-art dynamical forecast models, although out-of-sample predictions are less skillful, albeit over a small period. Forecasts over a longer out-of-sample period suggest there is true skill in the statistical models, comparable with that of dynamical forecasting models. They can be used both to help evaluate and to offer insight into the sources of predictability and limitations of dynamical models.

1. Introduction

The North Atlantic Oscillation (NAO) is a key element of Northern Hemisphere atmospheric circulation and is related to the storminess, wind speeds, surface air temperature, and precipitation variability over the North Atlantic Ocean and the adjacent continents of eastern North America and western Europe (e.g., Hurrell 1995; Hurrell et al. 2003). The NAO can be described as a seesaw of atmospheric mass between two nodes: a southern high pressure node over the subtropical Atlantic (Azores) and a northern low pressure node over Iceland. A positive NAO occurs with an increased pressure difference between the nodes, while a more negative NAO occurs as this difference decreases, although even for a negative NAO the absolute pressure difference is rarely reversed. This fluctuation of the pressure gradient between the nodes is directly proportional to changes in geostrophic wind speed. The NAO can be viewed as a consequence of storm track and

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/WAF-D-16-0124.s1.

Corresponding author: Richard Hall, rihall@lincoln.ac.uk

DOI: 10.1175/WAF-D-16-0124.1

© 2017 American Meteorological Society. For information regarding reuse of this content and general copyright information, consult the AMS Copyright Policy (www.ametsoc.org/PUBSReuseLicenses).

jet-stream variability, (e.g., Vallis and Gerber 2008), and there are significant correlations between jet-stream latitude and the NAO index (Woollings and Blackburn 2012). The NAO is most prominent in winter and explains up to one-third of the total variance in sea level pressure (SLP) over the North Atlantic (Hurrell and Deser 2009). It is highly variable, frequently changing phase over weeks and months, and there is little evidence for preferred time scales of variability (Hurrell and Deser 2009), with large variations from month to month, from year to year, and on decadal scales [see Hanna et al. (2015) for a recent review of NAO variations from 1899 to 2014]. Using daily data, Feldstein (2000) found NAO evolution to be consistent with a stochastic first-order autoregressive process with a time scale of around 10 days. However, Keeley et al. (2009) find up to 70% of winter NAO interannual variability is unexplained by short timescale variability and may therefore be externally forced.

There has been considerable debate over the extent to which the NAO is 1) driven by external climate factors and 2) is generated by internal atmospheric variability. For example, James and James (1989) report a long-term mode based on nonlinear feedbacks in the atmosphere creating low-frequency variability similar to the NAO. However, the NAO is not a consequence of local dynamics alone, as the storm-track pattern exists as a result of topographic forcing by the Rocky Mountains and the temperature contrast between the cold American continent and the warm Atlantic Ocean (Vallis and Gerber 2008). Furthermore, the enhanced interannual variability and positive trend in the NAO observed in the latter part of the twentieth century are greater than would be expected from internal atmospheric variability (Feldstein 2002) and are indicative of some external forcing such as from the ocean or sea ice (Hurrell and Deser 2009) that may not be properly reproduced in climate models (Scaife et al. 2009).

While some dynamical models exhibit only limited predictability in extratropical regions (e.g., Kim et al. 2012; Arribas et al. 2011; Jung et al. 2011), more recent work indicates there is likely to be a useful degree of predictability in the winter NAO. Folland et al. (2012) use a regression approach to forecast European winter temperatures based on a range of predictors, and recent work with dynamical forecast models (Riddle et al. 2013; Scaife et al. 2014) concludes that important aspects of winter climate and the NAO are predictable months ahead, with a high proportion of the variance being accounted for by the models (Scaife et al. 2014). A number of potential predictors have been identified: El Niño–Southern Oscillation (ENSO; e.g., Bell et al. 2009), spring North Atlantic sea surface temperatures (SSTs; e.g., Rodwell and Folland 2002), tropical volcanic eruptions (e.g., Robock and Mao 1995), Arctic sea ice extent (e.g., Strong and Magnusdottir 2011), the stratospheric quasi-biennial oscillation (QBO; Ebdon 1975), and autumn Eurasian snow cover (e.g., Cohen and Jones 2011) have all been linked with North Atlantic atmospheric circulation variability (Hall et al. 2015). Links have also been suggested between tropical SST anomalies and extratropical seasonal variability (e.g., Bader and Latif 2003; Hoerling et al. 2004; Li et al. 2010), where the upward trend in the NAO from 1950 to 1999 is attributed to increased SST over the Indian Ocean. However, the magnitude of the observed change in the NAO was much greater in the observations than in atmospheric climate models (Scaife et al. 2009). An influence of solar variability on the winter NAO has also been identified (e.g., Ineson et al. 2011). Some success has been found when using some of these predictors to make seasonal forecasts of winter weather in the North Atlantic region [e.g., Riddle et al. (2013) for Eurasian snow cover; Folland et al. (2012) for QBO, volcanic eruptions, El Niño, and Atlantic SSTs]. However, the sources of predictability in dynamical models are largely unknown. Here, we use a simple NAO index to examine this range of potential predictors, and compare our results with the Met Office (UKMO) Global Seasonal Forecasting System 5 (GloSea5), which has high ocean resolution (0.25°) and 3-hourly atmosphere–ocean coupling, as well as a fully resolved stratosphere and interactive sea ice physics package (MacLachlan et al. 2014). While the coupled dynamical model is state of the art, a simple probabilistic approach based on regression methods may help to illuminate particular weaknesses or limitations in the dynamical models and help to identify sources of predictability. The focus is on forecasting the sign of the winter NAO, and while many studies have looked at individual predictors, here we include a wide range of explanatory variables.

2. Data

The UKMO construct their NAO index by subtracting the raw values of SLP for the Azores and Iceland, then normalizing, rather than (as is more typically done) normalizing the station data separately and then subtracting. However, this has little impact on the sign of winter mean anomalies. Here, we construct a simple NAO index using station data for Reykjavik, Iceland, and Ponta Delgada, in the Azores, supplied by Adam Phillips at NCAR, for the period 1956–2016, using the UKMO approach. The NAO index is normalized to the period 1993–2012, in accordance with Scaife et al. (2014) to allow comparison with GloSea5 data. Normalizing by 1981–2010 has no effect on the sign of the NAO for any

of the years in question. Winter is December–February (DJF); the year of the winter is given by the year of the January and February.

A range of potential predictors is examined, chosen initially on the basis of a review of the literature. For the ENSO, a normalized Niño-3.4 index (N3.4) is used from 1956 to 2016, based on SST from HadISST1 (Rayner et al. 2003). A nonlinear relationship between ENSO events and the Atlantic sector has previously been observed, whereby moderate El Niño events are related to a negative winter NAO, whereas stronger events, with stronger SST anomalies in the eastern Pacific (greater than 1.5°C) do not produce an NAO-like response (Toniazzo and Scaife 2006; Bell et al. 2009). For example, in 2015/16 a strong El Niño did not produce the negative NAO response that would be anticipated from a moderate event. Therefore, following Folland et al. (2012), a discontinuous El Niño index is also used, with N3.4 values less than ±1 standard deviation of their seasonal variability equating to 0, values more negative than −1 are set to −1, and values from +1 to +1.75 are set to 1, with values above +1.75 being again set to zero, to reflect the nonlinearity of the forcing of different SSTs. Both versions of the ENSO index are available for selection in the regression models but the selection of one precludes the inclusion of the other.

Two metrics of Atlantic SST are used for 1956–2016. Unsmoothed Atlantic multidecadal oscillation (AMO) data (Enfield et al. 2001) are obtained from the Earth System Research Laboratory (www.esrl.noaa.gov/psd/data/timeseries/AMO), based on the Kaplan SST dataset (Kaplan et al. 1998, updated). A North Atlantic SST tripole index is developed using the methodology of Czaja and Marshall (2001). It is the SST anomaly taken over 40°–55°N, 60°–40°W minus the anomaly over a southern box, 25°–35°N, 80°–60°W (see Fig. S1 in the online supplement to this paper). Anomalies are relative to the 1981–2010 climatology. This dipole lies to either side of the Gulf Stream, and the third southern node of the classic tripole mirrors the northern node identified here. A positive (negative) tripole index indicates higher (lower) positive SST anomalies in the northern sector compared with those in the southern sector and reflects a reduced (increased) temperature gradient between the two.

Tropical SSTs can affect the atmosphere through altered convective activity and divergence aloft, which can generate Rossby waves that propagate away from the source and are capable of influencing the extratropics (Hoskins and Karoly 1981). Tropical rainfall is used as a proxy for this tropical convection. Version 2 of the Global Precipitation Climatology Project provides global precipitation data at 2.5° resolution, based on satellite data for the period 1979–2016, at monthly resolution (Adler et al. 2003). Six subsections are taken from the tropics: three from the Pacific Ocean [west Pacific rainfall (WPR), 5°S–5°N, 120°–170°E; central Pacific rainfall (CPR), 5°S–5°N, 170°–220°E; and east Pacific rainfall (EPR), 5°S–5°N, 220°–270°E], two from the Indian Ocean [west Indian rainfall (WIR), 5°S–5°N, 50°–85°E; east Indian rainfall (EIR), 5°S–5°N, 85°–120°E], and one from the Atlantic Ocean [Atlantic rainfall (AR), 5°S–5°N, 0°–50°W]. These areas are shown in Fig. S1 and ensure coverage of all equatorial tropical oceans. To increase the number of predictors available from 1956, tropical SSTs for the regions above are taken from the HadISST1 dataset (Rayner et al. 2003), for use with the longer NAO time series only from 1956 onward, as an indicator of tropical convective activity.

The QBO is an oscillation of zonal equatorial stratospheric winds with a period of around 28 months. It has been shown to influence the strength of stratospheric polar vortex anomalies (Holton and Tan 1980; Anstey and Shepherd 2014), which can in turn propagate downward and impact upon the polar front jet stream and NAO, especially in the late winter (Baldwin and Dunkerton 2001). QBO data are obtained for 1956–2016 from the Free University of Berlin [www.geo.fu-berlin.de/met/ag/start/produkte/qbo/; Naujokat (1986, updated)]. The 30-hPa equatorial zonal wind speeds are used, following Hamilton (1984).

Solar cycle data are available in a variety of forms. Monthly sunspot numbers are used to create a normalized index (1956–2016), available from the Solar Influences Data Analysis Center (http://sidc.oma.be/). Regression is also carried out using a lead of 1–5 yr of the solar cycle over the NAO as recent studies suggest that there is a lagged North Atlantic climate response to solar variability (Scaife et al. 2013; Gray et al. 2013) in addition to a shorter time-scale response operating via changes in the stratospheric polar vortex (Ineson et al. 2011).

A volcanic index is derived according to Folland et al. (2012), which once again spans 1956–2016. The index is set to one for the two years following a tropical volcanic eruption, to allow for the lifetime of stratospheric volcanic aerosols, all other years being set to zero, with the years of volcanic eruptions being derived from Stenchikov et al. (2006). A positive NAO in winters following a major tropical eruption has been observed (e.g., Robock and Mao 1995).

Sea ice concentration data are taken from HadISST1 (Rayner et al. 2003) for the longer hindcast time series since 1956, while data from the National Snow and Ice Data Center (NSIDC) are used for hindcasts from 1980 (Cavalieri et al. 1996, updated). The correlation between the two datasets is very high for the period of overlap (1979–2014 November sea ice, $r = 0.98$). Data

are acquired for the whole of the Arctic, plus subregions identified as being of potential significance in the literature. Areas identified are the Barents–Kara Sea (BKI; 70°–85°N, 30°–90°E), NE Greenland (GI: 80°–90°N, 35°W–0°) and the area centered on the Laptev Sea (LVI; 70°–90°N, 60°–200°E), but including the east Siberian, Kara, and Chukchi Seas. Snow cover data for Eurasia (45°–80°N, 55°–150°E; 1979–2016) are obtained from Rutgers University; http://climate.rutgers.edu/snowcover/; Robinson et al. (2012)], to give the monthly snow cover extent. These regions are shown in Fig. S1.

The GloSea5 ensemble hindcast data with 24 ensemble members for 1993–2012 are supplied by the UKMO, together with operational forecasts for winters 2014–16 (the system was not operational in 2013). Operational forecast ensemble sizes vary in number: 31 in 2014 and 32 for 2015 and 2016.

All predictor datasets are normalized by subtracting the monthly mean and dividing by the monthly standard deviation for the period 1981–2010. Any trend in the data is retained. No tuning of the predictors is performed to obtain the initial statistical forecast models, although detrending of sea ice is used in a subsequent model.

3. Methods

a. Regression models

We use a simple multiple regression approach to identify linear aspects of predictability. Regression has already been shown to provide evidence of significant predictors of North Atlantic climate variability (Folland et al. 2012), and it is good scientific practice to start with a simple approach, which can then be further developed. Potential predictors have been identified for the winter NAO, based on the literature in section 1, and correlations between the various drivers and the NAO index at a range of monthly lead times, up to 1 yr ahead, with the exception of solar variability lead times, which were on a monthly basis for up to 5 yr ahead. The lead times selected in the models have the greatest explanatory power, although if this is similar for different lead times (a difference in R no greater than 0.02), the month chosen is that with the more plausible physical association, based on known relationships in the literature. Predictors for the multiple regression models are identified by forward selection (e.g., Wilks 2011), with synchronous drivers omitted. In sensitivity tests, other methods of selection had qualitatively very little impact upon the predictors selected. The stopping criterion is identified by calculating a t value, which is defined as the ratio of the regression coefficient estimate of each predictor to its standard error. Forward selection is continued until no further predictors can be added with $p \leq 0.05$. Predictors are not included in the model if they have a significant ($p \leq 0.05$) correlation with any of the prior selected predictors, to minimize multicollinearity. The Akaike information criterion (AIC) produced very similar results for predictor selection but was slightly more liberal with predictors of marginal significance.

Statistical hindcasts are constructed from 1980 to 2012, hereafter identified as N80, covering the mainstream satellite era, and from 1956 to 2012 (N56). In addition, a 20-yr hindcast (1993–2012; N93) is constructed for direct comparison with GloSea5. The hindcast time series are cross validated using leave-one-out cross validation, to ensure that the time series generated is not correlated with the year being predicted. Cross validation is also applied in the production of normalized predictor values. In principle, therefore, a separate model with different coefficients is created for each year.

b. Simple ensemble creation

The variance of the fit generated from each regression model is less than that of the observed time series. This is because the regression model captures some of the forced signal but not the unforced internal atmospheric variability. Observations should be statistically indistinguishable from the ensemble forecasts, so in order to generate a consistent ensemble, we incorporate an unforced noise component. The variance due to both noise and the part of the forced signal not captured by the model can be taken as

$$\text{Var(noise)} = \text{Var(obs)} - \text{Var(ensemble mean)}. \quad (1)$$

The noise is added to the ensemble mean by generating random numbers from a Gaussian distribution with standard deviation equal to $\sqrt{\text{Var(noise)}}$, in Eq. (1) and with a mean of zero. For a time series consisting of n years, n random numbers from the distribution are generated, and one of these values is added to each of the annual predicted values to generate an ensemble member; this process is then repeated for the required number of ensemble members. Note that this adjustment is only applied for the generation of ensemble members and does not affect the ensemble mean used in hindcasting, which is generated directly by the regression model. This simple method also assumes the same spread in each year. To compare with GloSea5 dynamical forecasting data, 24 ensemble members are created. The total variance of the 24-member ensemble is very close to the variance of the observed time series and is statistically indistinguishable. GloSea5 ensemble members are generated by the forecasting general circulation model and are averaged to create the GloSea5 ensemble mean.

Verification rank histograms (VRHs) are plotted to establish the extent to which the observed time series differs from the ensemble members (both GloSea5 and statistical models). These indicate whether the hindcast ensembles include the observations as equally likely members (e.g., Wilks 2011). It is important to distinguish between an uneven distribution due to sampling variations and true deviations from a uniform distribution. An alternative to the chi-square goodness-of-fit test is to use nonparametric tests from the Cramér–von Mises group of tests, specifically the Watson (Watson 1961) and Anderson–Darling (Anderson and Darling 1952) statistics. These methods have been developed for discrete distributions by Choulakian et al. (1994). Both tests are used, as the Watson test has been found to be more sensitive to U-shaped or peaked distributions, while the Anderson–Darling test is more sensitive to bias or rank (Elmore 2005).

c. Probabilistic hindcasts

The ensemble mean as derived above gives a deterministic NAO forecast. We also present a probabilistic hindcast of the sign of the NAO. We choose to use this hindcast threshold (NAO $\leq$ 0) as the number of forecast–observation pairs will be increased compared with a more extreme threshold, for example of the NAO being less than -1, which will have relatively few occurrences in the observed record. Actual occurrences of the observed NAO at or below an NAO index value of zero are expressed in binary form (1 = occurs, 0 = does not occur) for each year. Probabilistic hindcasts are constructed from the 24-member ensemble. Probability is calculated as the proportion of the 24 members giving predicted NAO values at or below zero for each year. As the ensemble size is not large, a simple adjustment is made for small sample size (Wilks 2006), such that the probability of the forecast f being less than or equal to a given quantile q (in this case NAO $\leq$ 0) is

$$\Pr(f \leq q) = \frac{\text{Rank}(q) - 1/3}{(n_{\text{ens}} + 1) + 1/3} \tag{2}$$

or, in the case of the positive forecasts, it is

$$\Pr(f \geq q) = 1 - \frac{\text{Rank}(q) - 1/3}{(n_{\text{ens}} + 1) + 1/3}, \tag{3}$$

where Rank(q) shows the rank of the quantile in question in terms of its position within the ensemble forecast for a given year. Here, Rank(q) = 1 if it is smaller than all n_{ens} ensemble members and Rank(q) = n_{ens} + 1 if it is larger than all members. The further adjustments in the equation ensure that the value obtained is approximately equal to the median of the estimated sampling distribution of the cumulative probability in question.

d. Probabilistic forecast verification

A wide range of forecast verification tools can be used. Here, the Brier score (BS), Brier skill score (BSS), reliability diagrams, and relative operating characteristic (ROC) diagrams are used to provide a range of metrics for assessing the forecast (e.g., Wilks 2011). Consistency bars (Bröcker and Smith 2007) are added to the reliability diagrams, which give an indication of how far the observed relative frequency is likely to depart from the diagonal if the forecast is perfectly reliable. Bars are shown for the 95% confidence limits.

Ten forecast probability bins are used for the initial analysis, although the sensitivity of the results to bin size is addressed by rerunning the verification tests for five bins. Five bins provides a more optimal representation for the reliability diagram and verification statistics, giving sufficient bins while ensuring these bins are populated, as well as reducing the noise evident in the initial 10-bin run.

The area under the ROC curve (ROC area) can be tested for significance against the null hypothesis that the area equals 0.5. The ROC area is equivalent to the Mann–Whitney U statistic testing forecast probabilities for cases when the forecast occurred compared with occasions when events did not occur (Mason and Graham 2002).

VERIFICATION BY FORECASTING FUTURE NAO VALUES

The ability of the N56, N80, and N93 models to provide genuine forecasts of the NAO is initially tested on the years 2013–16, which are years outside the period over which the model is developed (the training period). However, as this is a very small sample, increasing incrementally by one value each year, it is necessary to also use an alternative approach, which is based on a larger out-of-sample group of years. Therefore, a regression model is developed based on the training period 1980–97 and then tested on another period (the testing period) of similar length (1998–2016). The NAO value for each year in the testing period is predicted using values of the selected predictors in the regression equation. Statistical models are frequently overtuned as predictors are often based upon those identified from observational associations, and so could be a consequence of noise rather than a meaningful physical

TABLE 1. Regression coefficients of predictors selected for the regression models N56, N80, and N93. The *y*-intercept term is *A*, and R^2 and cross-validated R^2 (xvR^2) values are given. Columns indicate the following: OctN3.4 = October N3.4 discontinuous index, SepWISST = September west Indian Ocean tropical SSTs (used in N56 only), Feb 2yr lead SS = February solar activity at a lead of 2 yr (34 months total), NovBKI = November Barents–Kara Sea ice (NSIDC), OctWIR = October west Indian Ocean tropical rainfall (only available for N80 and N93), JunTRI = June Atlantic tripole SST, and OctAMO = October AMO. Within the table, NA denotes a predictor is not available for a particular model. All R^2 values are significant at $p \leq 0.05$, through calculation of the *F* statistic. Significance values for predictor coefficients are set within parentheses below each coefficient.

Model	*A*	Oct N.34	Sep WISST	Feb 2yr lead SS	Nov BKI	Oct WIR	Jun tripole	Oct AMO	R^2	xvR^2
N56	−0.004	−0.91 (3×10^{-4})	0.24 (1×10^{-3})	0.28 (0.02)	—	NA	—	—	0.34	0.24
N80	0.01	−0.79 (6×10^{-4})	—	—	0.43 (4×10^{-2})	0.38 (3×10^{-4})	0.15 (0.05)	—	0.68	0.58
N93	0.15	−0.97 (9×10^{-4})	—	—	0.34 (4×10^{-2})	0.51 (2×10^{-4})	—	−1.75 (0.01)	0.78	0.63

connection. Use of a testing period assists in separating the noise and coincidental relationships from the physical connections.

4. Results

a. Deterministic hindcasts

In this section we present the regression models developed for each time series and illustrate their performance as deterministic hindcasts. Table 1 shows the regression coefficients of the predictors selected for the models. The R^2 values and the *y*-intercept term *A* are given, allowing straightforward construction of the regression equations.

The models, although differing in some aspects of predictor selection, identify similar potential predictors of the winter NAO. The models demonstrate predominantly tropical and Arctic influences. October N3.4 is present in all models while a sea ice term is also used in N80 and N93 (November BKI). Tropical influences from the western Indian Ocean are also represented in all models by October precipitation (N80 and N93) and September SST in N56. In addition, N80 shows an extratropical influence from the June North Atlantic tripole, while in N93 the October AMO is selected as a predictor and solar forcing is also significant in the longer time series. The solar forcing term here is at a lead of around 3 yr, consistent with that identified from other studies (Scaife et al. 2013; Gray et al. 2013; Andrews et al. 2015). A number of very similar correlations are found at lead times ranging from 6 months to 3 yr. Extending models back to 1956 provides no additional skill. The R^2 values are higher for the models based on post-1980 data, perhaps reflecting the improvement in observational data quality during the satellite era, although in the models developed, all predictors identified are available for both longer and shorter series, if tropical SST is substituted for tropical precipitation in the longer series. It could also be that the early period is less predictable.

Figure 1 shows the observed NAO index together with the time series of predicted NAO values derived from the models above. It is clearly seen that for N56, the correlation between observations and predictions is less good before 1979 ($r = 0.33$) and insignificant ($p \geq 0.05$) compared with the post-1979 period ($r = 0.48, p \leq 0.05$). This is likely to be at least partly due to the improved data quality of predictors such as sea ice post-1979 as a result of the availability of satellite data. There are periods where all models show a close match with observations (e.g., 2008–12) while during other periods there is greater divergence (e.g., 2001–05). This possible variability in predictive skill on decadal scales breaks the assumption of stationarity.

An indication of the uncertainty of hindcasts for individual years is obtained by identifying years when the observed value lies outside the hindcast 95% confidence limits defined by ±1.96 times the ensemble noise standard deviation [1.96SD; Eq. (1)], shown in Table 2.

The observed years lying outside the 1.96SD range of the ensemble mean can be identified as poorly predicted. The number of these cases is small for each hindcast (Table 2). It would be expected that 1 year in 20 (5%) would be outside the 1.96SD range by chance alone. For N56, 9% of years are outside this range (five years), while N93 and GloSea5 have one more year identified than would be expected by chance (10%), and N80 has 12% of years outside the range (four years), but sample sizes are small so results may still be due to chance. The year 1996 is consistently poorly hindcast for all models except N56, a positive NAO being predicted in every case while a negative NAO was observed. The year 2012 is also poorly hindcast, being underpredicted by N80 and N93. N56 manages to predict well years that

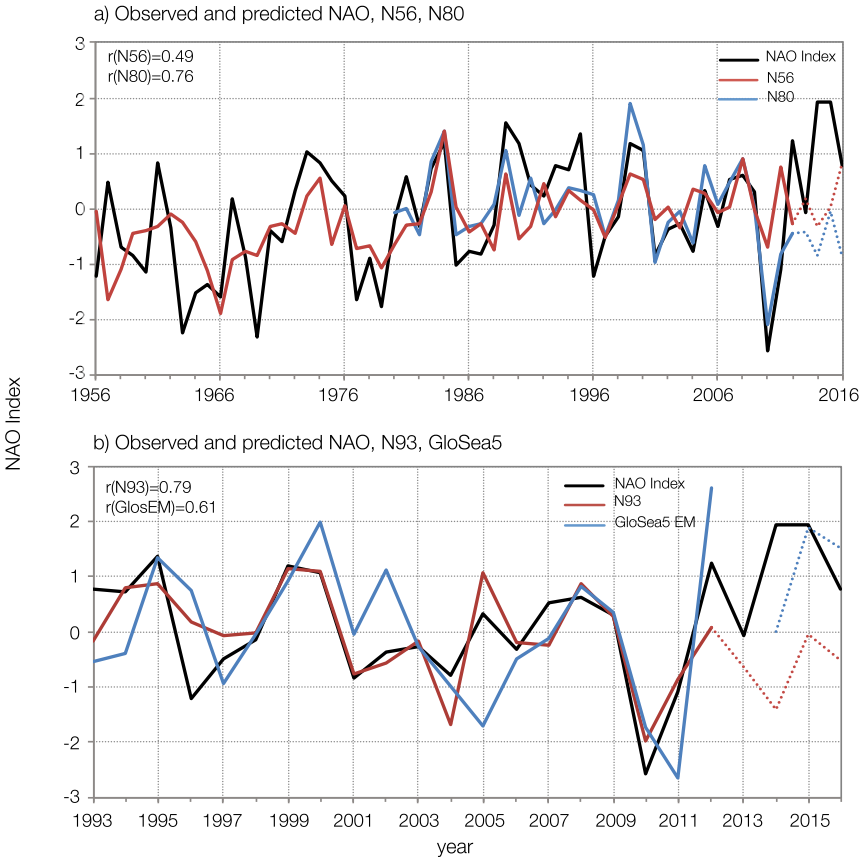

FIG. 1. (a) Observed (black solid) and predicted cross-validated NAO time series (N56, red solid; N80, blue solid), based on the statistical models. (b) As in (a), but for GloSea5 (blue solid) and N93 (red solid) compared with the observed NAO index (black solid). Out-of-sample forecasts are shown as dotted lines. Note the different time scales along the axes.

are poorly predicted by other models, but in turn predicts different years poorly. Such variations between poor hindcast years in different models, while in part attributable to random fluctuations, can give insights into possible reasons for the poor performance in a particular year. For example, 2005 is a poor hindcast in GloSea5 with a negative predicted (−0.53) and a positive observed NAO (0.78). The correct positive hindcasts in the statistical models for this year can in part be related to a strong positive signal from the October west Indian Ocean precipitation (N80 and N93) or the September west Indian Ocean SST value (N56). However, the absence of a sea ice term in N56 means that 2010 is poorly predicted in this model. Although the sign is correct, the predicted negative NAO is far too weak. It is likely that the increasing negative sea ice trend in the autumn accounts for the underprediction of 2012 in N80 and N93 (see below).

The year 1996 was an important one, marking the end of the positive NAO trend of the late twentieth century and coinciding with a rapid warming of the North Atlantic subpolar gyre (Robson et al. 2012), which seems not to be evident in the resulting predictions, even though the June tripole is used as a predictor in N80 and models are able to predict this event (Hermanson et al. 2014).

In contrast, winter 2011 is well predicted by N93 and N80. For this year the Atlantic SST signal outweighed the sea ice signal in both forecasts. For N93, the October AMO provides 28.8% of the negative NAO forcing, compared with 16.6% from sea ice, while for N80, sea ice and the June Atlantic SST tripole provided 24.6% and

TABLE 2. Years for which the difference between forecast–observation pairs is greater than 1.96SD ensemble noise for the year in question.

Forecast	Years
N56	1957, 1963, 1990, 2010, 2011
N80	1990, 1995, 1996, 2012
N93	1996, 2012
GloSea 5 index	1996, 2005

28% of negative forcing, respectively. This is consistent with the conclusions of Maidens et al. (2013), who found that Atlantic SSTs were a major contributory factor to the negative NAO of this year. Interestingly though, in both models around 50% of the negative forcing comes from west Indian Ocean rainfall, which is indicative of convective activity and divergence leading to Rossby wave propagation. Such a possibility is indeed explicitly acknowledged by Maidens et al. (2013).

b. *Ensemble predictions*

We now present the ensemble hindcast values and evaluate the effectiveness of the ensemble by comparing the observed NAO values to those of the ensemble members, using VRHs. Figure 2 shows ensemble predictions (gray dots) compared with the observed NAO and ensemble mean (the predicted values), together with the VRH constructed from the ensemble.

At first sight, the VRHs appear uneven (Fig. 2) and difficult to interpret, but this could be due to the relatively small ensemble size resulting in statistical noise at certain ranks. There is no discernible systematic bias in the histograms for the statistical forecast models, and the Watson and Anderson–Darling statistics suggest that the null hypothesis of a uniform distribution cannot be rejected at $p \leq 0.05$ for any of the ensembles. The statistics are similarly inconclusive for VRHs using raw pressure differences between the Azores and Iceland (not shown). Other statistics of the GloSea5 ensemble (Eade et al. 2014) do however confirm the statistical significance of overdispersion in GloSea5 when using raw rather than standardized data. The small sample sizes do not allow accurate identification of systematic bias among ensemble members.

c. *Probabilistic hindcast verification*

This section examines the probabilistic forecasts outlined in section 3c. Table S1 summarizes the probabilistic hindcasts from statistical models and GloSea5 and the observed NAO and these data are used as the basis for probabilistic forecast verification. Forecast verification statistics for the NAO ≤ 0 probabilistic forecasts are presented in Table 3.

All models show positive skill, and the scores from N80 are usually better than GloSea5 in terms of accuracy (BS) and skill relative to climatology (BSS) (Table 3). However, some of this may be a consequence of an increased length of time covered, although N93 performs even better, over the same time period as GloSea5. N56 generally has the poorest set of verification scores, which would be expected as the model is a less good fit to the observations (Fig. 1a), although the scores are comparable to those of GloSea5. The correlation skill of N56 over the years 1980–2012 is 0.41, compared to 0.76 for N80.

Figure 3 presents reliability diagrams for the statistical probabilistic hindcasts and for GloSea5, for the probability of the NAO being less than or equal to zero, based on five bins for probability forecasts. It will be noted that the consistency bars are wide, a consequence of the small sample sizes. All points plotted on the curve lie within the consistency bars, but are on occasion at the extreme ends of the bar, when the number of forecasts within a probability bin is low, which is again a consequence of small sample size.

It is difficult to compare reliability diagrams between models because of the degree of fluctuation due to small sample size. All diagrams have a positive slope indicating that as forecast probability increases, so does the frequency with which the event is observed; therefore, all are to some extent reliable. Refinement distributions, as shown by the inset histograms, indicate that all forecasts show some sharpness, in that all forecast probability bins are used. The sharper forecasts are N80, N93, and GloSea5, where there are more instances of extreme high or low probabilities, rather than clustering around climatological probability values.

The greatest departures from the diagonal occur when there are few occurrences within a forecast probability bin. In such cases, one further occurrence will make a large difference to the proximity of the curve to the diagonal. Normalized values of GloSea5 show greater reliability than the raw pressure differences (not shown), indicating that reliability can be added to a forecast by data processing techniques.

ROC areas (Table 3) show the forecasts to yield good discrimination between events and nonevents and to be potentially useful. N80 and N93 have the highest ROC area scores although values for all statistical models and GloSea5 are high and statistically significant ($p \leq 0.05$).

N80 and N93 provide the best probabilistic forecast models in terms of skill, reliability, resolution, and accuracy and compare well with GloSea5, although it is hard to say that one forecast is better than another due to small sample sizes. Attempting to use a longer time series does not necessarily produce a better-quality forecast. With N56 this is likely to be due to reduced data quality in the presatellite era or, perhaps, due to a change in inherent predictability.

d. *Using the models for out-of-sample forecasting*

Here, we apply the regression models outlined in section 4a to out-of-sample forecasting for the years 2013–16. This acts as a better indicator of the models' true forecasting potential. A model developed using only data from the years 1980–97 is then applied to

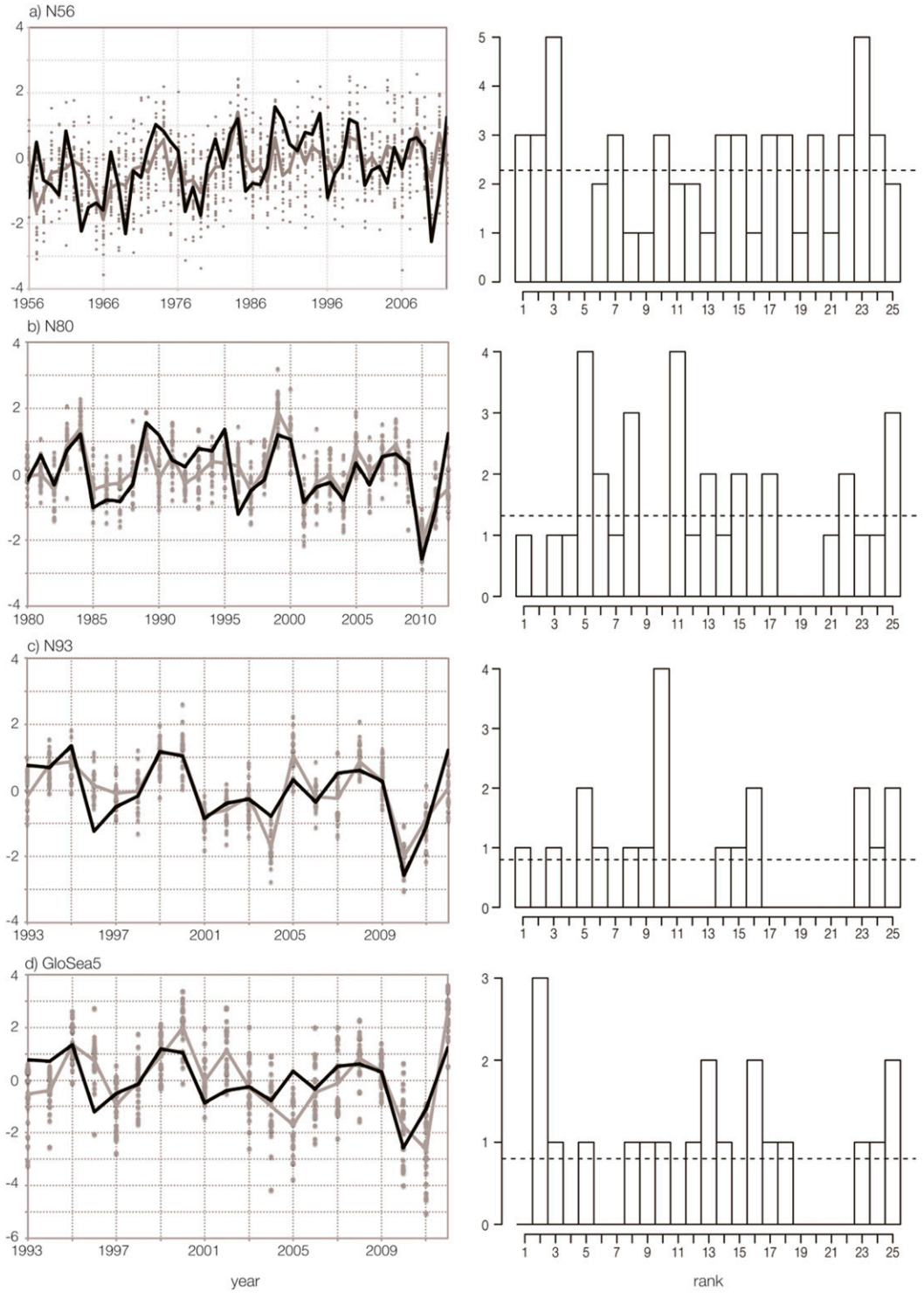

FIG. 2. Ensemble members (gray dots), ensemble mean (gray line), and observed NAO values (boldface black line) together with VRHs for (a) N56, (b) N80, (c) N93, and (d) GloSea5. Dashed lines in histograms indicate expected values of counts for each rank if the observations are equiprobable at all ranks.

TABLE 3. Verification statistics for probability forecasts using five bins. All ROC area values are considered significant ($p \leq 0.05$).

NAO ≤ 0	BS	BSS	ROC area
N80	0.15	0.40	0.87
N93	0.09	0.64	0.96
GloSea5 index	0.21	0.18	0.76
N56	0.20	0.17	0.71

forecast the winter NAO for 1998–2016, giving a longer period of forecasting.

Forecast values for the years 2013–16 based on the statistical models are shown in Table 4. Here, the statistical forecasts appear to be less well matched to observations than for the period 1980–2012. N80 and N93 issue four negative forecasts each, and only one of these is matched in sign by observations (2013). N56 issues one negative and three positive forecasts, and the sign of the observed NAO is correct for 2015 and 2016 (positive). Nine out of the 12 forecasts have a probability greater than or equal to 0.5 of a negative NAO occurring, although 7 of these are for N80 and N93, out of a total of 8 forecasts. However, plotting the results reveals that the predicted values for N93 and N80 track the observed values for most years but with a systematic negative bias (Fig. 1). The year 2014 is an exception to this, appearing as a relative minimum in all models including GloSea5, while the observed NAO is a relative maximum. Differences between forecast and observed values for 2013 do not appear to be distinctly different in magnitude

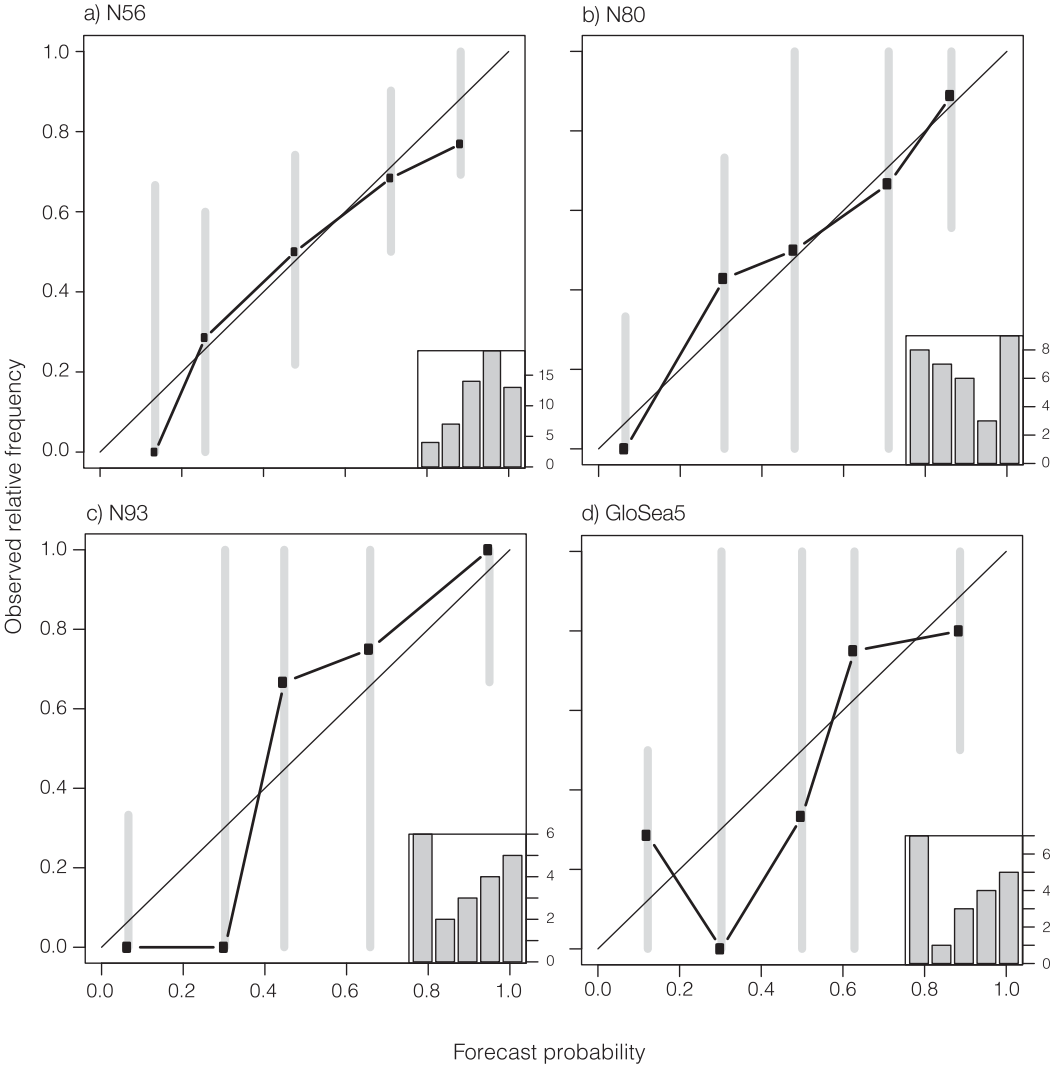

FIG. 3. Reliability diagrams for forecast models NAO ≤ 0, for five bins. Histograms in bottom-right-hand corner show the frequency of occurrence for each forecast probability bin. Gray vertical lines are the consistency bars for the 95% confidence interval.

TABLE 4. Observed and forecast values for the years 2013–16 from the statistical models. Observed and ensemble mean (forecast) NAO values and probabilistic forecasts are given. Boldface values in the forecasts column show that the sign of the NAO is predicted correctly for the year in question.

		N56		N80		N93		GloSea5	
Year	Observed NAO	Forecast NAO	Pr(NAO ≤ 0)	Forecast NAO	Pr(NAO ≤ 0)	Forecast NAO	Pr(NAO ≤ 0)	Forecast NAO	Pr(NAO ≤ 0)
2013	−0.06	0.19	0.42	**−0.42**	0.74	**−0.65**	0.70	NA	NA
2014	1.93	−0.31	0.62	−0.84	0.89	−1.41	0.97	**0.01**	0.48
2015	1.93	**0.03**	0.50	−0.05	0.62	−0.05	0.38	**1.89**	0.02
2016	0.77	**0.83**	0.08	−0.87	0.97	−0.52	0.93	**1.52**	0.05

from those seen during 1980–2012; however, the differences for 2014–16 are large relative to differences over 1980–2012. GloSea5 forecasts all predict a positive NAO (2014–16), which matches the observed NAO, although in 2014 the prediction is only just positive (0.01; see Table 4). While the 2015 forecast has a very close match to the observed value, 2014 was underestimated and 2016 overestimated by GloSea5. Therefore, the N80 and N93 statistical models show a systematic negative bias for out-of-sample forecasting while matching relative maxima and minima with those of observations, whereas N56 does not show the negative bias but the match of maxima and minima is less good. GloSea5 manages to successfully predict the sign of the winter NAO and matches the interannual change of magnitude of the NAO with the exception of 2014. Forecast skill for GloSea5 in winter 2016 is likely to come from ENSO and the QBO (Scaife et al. 2017). The strength of the positive winter NAO in 2014 is underpredicted in all models including GloSea5, suggesting a greater role for internal variability or a factor not well represented in any of the models.

The forecast model developed for the training period 1980–97 is only based on two predictors, November Barents–Kara Sea ice and the October N3.4 adjusted index using the selection criteria outlined above for identifying predictors over the training period:

$$\mathrm{DJFNAO} = -0.14 + 0.71\mathrm{NovBKI} - 0.74\mathrm{OctN3.4}$$

$$R^2 = 0.56. \tag{4}$$

Figure 4 shows the fit during the training period (1980–97) together with the subsequent fit of forecasts during the testing period (1998–2016). The correlation between observed and forecast NAO results for the training period is significant (0.75, $p \leq 0.05$) while that for the verification period is 0.32 (not significant; $p \geq 0.05$). However, for most of the testing period, the match is

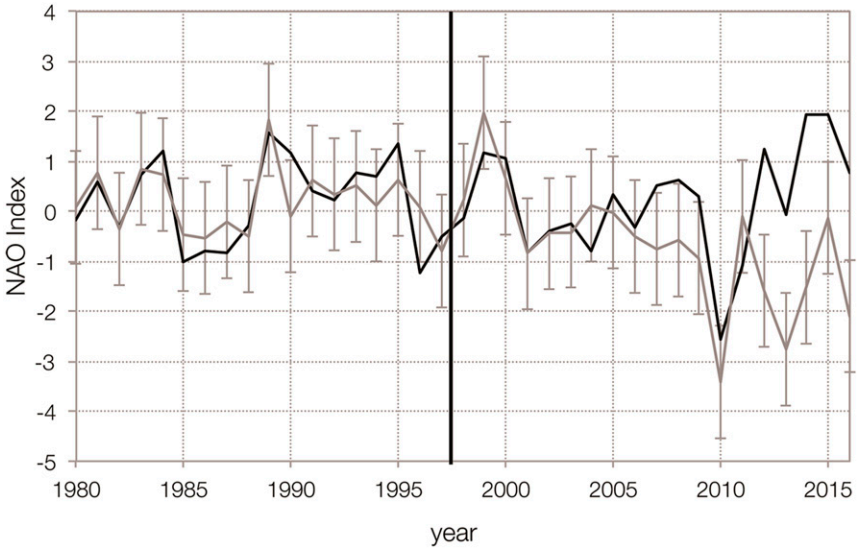

FIG. 4. Observed NAO (black) and predicted NAO values (gray) for the testing period 1998–2015, based on a training period model covering 1980–97. Black vertical line denotes the end of the training period and the start of the testing period. Error bars are for ±1.96 ensemble noise standard deviation for each year.

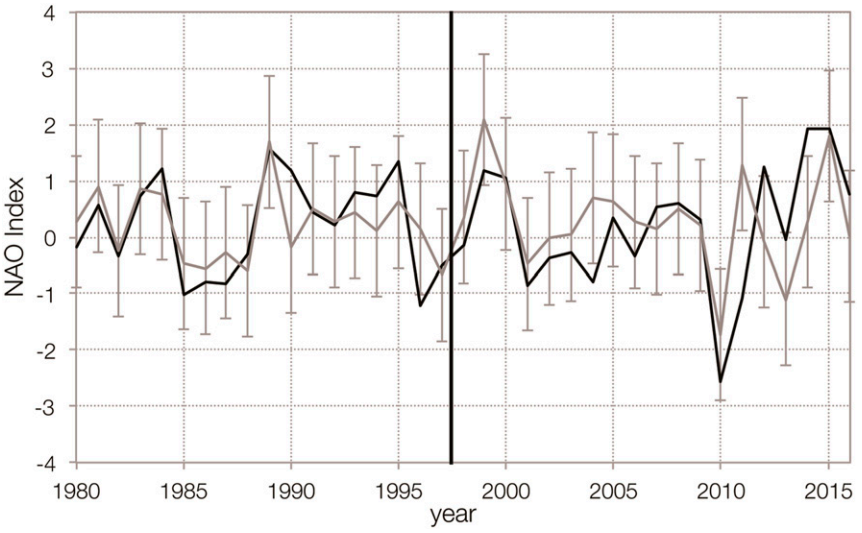

FIG. 5. As in Fig. 4, but using detrended sea ice data.

significant (1998–2011; $r = 0.75$). The model forecasts also appear to reproduce the increased variability present in the NAO during the testing period and replicate the magnitude of extreme NAO events, such as winter 2010. It is in the last 5 yr that the observations and forecasts become less well correlated, with forecasts being too negative, consistent with the results from statistical forecasts in Table 4. As there is no input from the N3.4 predictor for the years 2011–16, and with the index being set to zero for these years, the negative bias in the forecast NAO must come from the sea ice. Both forecasts for the testing period and the autumn sea ice extent show a negative trend, while the observed NAO does not. Winter 2013 was preceded by a very low Barents–Kara sea ice value (November 2012), resulting in a strong predicted negative NAO of the same order of magnitude as that for 2009–10, but this is not reflected in the observed winter NAO in Fig. 4, where the dip is relatively slight. In reality, the very low sea ice value appears to either be offset by other drivers not included in the model or the sea ice is given too much weight in the statistical method. The sea ice recovered in 2013 and 2014, which is reflected in the models and observations of the NAO for 2014 and 2015, but there is still an underestimation of the forecast compared with the observed index. This may be due to the influence of the sea ice trend, which is quadratic over the period, steepening since 2000 (Fig. S2). For the testing period, observations outside the forecast error bars occur in 1996, 2007–09, and all years after 2011. In all cases observations are more positive than forecast NAO values, confirming the systematic negative bias evident in forecasts from the latter period, although relative maxima and minima in forecasts and observations frequently coincide. The influence of the sea ice trend is supported by out-of-sample forecasts for N56, which does not contain a sea ice term and shows no negative bias.

A further statistical model for the training period 1980–97 is therefore constructed using detrended sea ice data, to remove the influence of the sea ice trend on forecast NAO values. The same predictors are selected, but with slightly different coefficients and a positive y-intercept term and with an R^2 value that is very similar:

$$\mathrm{DJFNAO} = 0.20 + 0.70\mathrm{NovBKI(detrended)} - 0.76\mathrm{OctN3.4} \quad R^2 = 0.53. \tag{5}$$

When the forecasts are made for the testing period (1998–2016) using this model, and the systematic bias in statistical forecasts is eliminated (Fig. 5), a much closer match is obtained between forecast–observation pairs, with only three of the six most recent observations lying outside the forecast error bars. A higher NAO value is predicted in 2011 while 2012 and 2014 predict lower NAO values than observed. Correlations between observed and predicted values are now 0.73 for the training period and 0.56 for the testing period, both of which are significant ($p \leq 0.05$). When detrended sea ice data are used, the model in Eq. (2) correctly predicts the sign of the NAO in 13 out of 19 yr for the testing period, compared with only 9 when the trend is retained. It appears that interannual variability of sea ice is a better predictor of the winter NAO than absolute sea ice values and that inclusion of the sea ice trend leads to an overestimate of the influence of sea ice. The correlation skill of 0.56 for

TABLE 5. Verification statistics for 1980–97 regression model, with the sea ice trend retained and removed, for the training period (1980–97) and the testing period (1998–2015).

NAO forecast	BS	BSS	ROC area
Trend-in sea ice			
1980–97	0.10	0.60	0.94[a]
1998–2016	0.32	−0.27	0.54
Overall	0.21	0.17	0.73[a]
Detrended sea ice			
1980–1997	0.12	0.51	0.94[a]
1998–2016	0.22	0.11	0.70
Overall	0.17	0.31	0.80[a]

[a] ROC values are considered significant ($p \leq 0.05$).

the testing period compares with the correlation of 0.61 achieved by GloSea5.

Verification data for the 1980–97 trend-in model confirm that NAO forecasts for 1998–2016 for this model have little skill and accuracy (Table 5), which improve considerably if the sea ice trend is removed. Overall for the whole time period (1980–2016), removing the sea ice trend improves the verification statistics (improved BSS, a small improvement in BS).

5. Discussion

While much work has suggested that the variability of the NAO/Arctic Oscillation (AO) is due to internal atmospheric dynamics (e.g., James and James 1989; Hurrell et al. 2003), analysis with GloSea5 and statistical models indicates that there does appear to be a significant predictable component in the winter NAO, derived from slowly varying boundary conditions. It is possible to produce statistical hindcasts for the NAO that have high levels of skill and reliability. However, although care has been taken not to overfit the regression models with too many predictors, it is still possible that these models are overtuned, as they have fared more poorly in recent out-of-sample years, compared with the dynamical forecasts of GloSea5. Associations with potential predictors could be nonstationary, or simply a result of noise, and therefore not necessarily indicative of true relationships. Nevertheless, such models may help to provide a benchmark for dynamical models, although it must be borne in mind that they rely upon chosen predictors that follow data inspection through observational studies. There is reasonable success in testing regression models against independent verification data, shown by the ability of models to forecast the NAO for 2013–16, matching fluctuations in the observed NAO albeit with an apparent negative bias.

The June tripole has a 6-month lead time over the winter NAO and a mechanism has been established that makes this link. The late spring/early summer tripole pattern is preserved beneath the summer thermocline. The thermocline breaks down in winter, allowing the tripole signal to reemerge (Rodwell et al. 1999; Deser et al. 2003).

Regarding the association with west Indian Ocean rainfall and SSTs, the time scale for Rossby wave propagation to midlatitudes of 1–2 weeks (Hoskins and Karoly 1981) does not appear to match the lead time found here of 2–3 months prior to the start of winter. However, Li et al. (2010) report that an annular mode response to tropical Indian Ocean heating is not achieved until after around 45 days. Although the Rossby wave propagation time scale is of the order of 2 weeks, this does not produce an annular mode response, which is dependent on the presence of feedback from transient eddies onto the large-scale atmospheric flow. This suggests a plausible physical mechanism for the time scales found here, with persistent patterns extending the time scale, although the potential mechanism should be further investigated.

A solar variability term with a lead time of 34 months is used in N56, as a number of significant correlations were identified with lead times ranging from 6 months to 2 yr. Models produced using these different lead times for solar variability are qualitatively very similar, with the same predictors, and very similar coefficients and R^2 values. This lagged response of the NAO to solar variability has been shown both in observations (e.g., Scaife et al. 2013) and model experiments (e.g., Andrews et al. 2015). The mechanism suggested is that the North Atlantic upper-ocean temperatures provide memory of the solar variability, which produces a lagged NAO response (Andrews et al. 2015).

The testing of a model over a longer verification period showed that the ensemble mean forecasts were frequently able to capture the sign and magnitude of the observed NAO (Fig. 4). However, forecasts for recent years (since 2007) show a negative bias, which is very strongly evident for the 1980–97-based model, where the only predictors are November Barents–Kara Sea ice and October N3.4 (Fig. 4), although relative maxima and minima are reasonably well reproduced. This suggests that the negative bias comes from the marked decline in sea ice as the N3.4 index was set to zero for most of these years, and the statistical models therefore overestimate the influence of sea ice. This results in negative forecasts being issued too frequently and poorer skill in these negative forecasts. This is evident to some extent in all statistical models with the exception of N56, which contains no sea ice term but is particularly noticeable in the very recent years since 2007. It has been demonstrated that removing the sea ice trend can improve the

accuracy of these forecasts and remove the systematic negative bias; therefore, the interannual variability of sea ice may be a better predictor of the winter NAO rather than absolute sea ice values (Fig. 5; Table 3).

The sea ice signal contains two components. First the trend, up to 60% of which is a likely consequence of greenhouse gas forcing and increased Arctic amplification, with the remainder being due to internal variability (Kay et al. 2011; Stroeve et al. 2012). Second, there is the interannual variability of autumn sea ice coverage, which can be influenced by initial ice conditions, summer weather conditions, and storms in the Arctic, which can affect the amount of solar radiation absorbed by the ocean, and poleward atmospheric moisture and heat fluxes and oceanic heat fluxes (Holland et al. 2011; Park et al. 2015; Stroeve et al. 2016). It is possible that some of these factors influencing sea ice interannual variability, such as shifts in the Gulf Stream, may themselves be effective predictors of the NAO, and the sea ice in fact modulates such a signal (e.g., Sato et al. 2014). Furthermore, using sea ice data with the trend retained incorporates a greenhouse gas forcing signal that, while influencing the sea ice negative trend, at the same time is expected to lead to a mean northward shift in the jet stream, as a result of warming in the tropical upper troposphere (e.g., Butler et al. 2010). The fact that this is not seen in the Atlantic sector (e.g., Barnes and Polvani 2015) could be a result of greenhouse forcing on the jet stream and Arctic amplification effectively working in opposite directions (e.g., Barnes and Screen 2015), with little net change in jet latitude seen. Thus, if the sea ice trend is retained, the compensating effect of greenhouse forcing through tropical warming is not accounted for in the models, hence the systematic negative bias.

It is therefore recommended that for future forecasts, a detrended sea ice index be used, but also there is scope for retaining the sea ice trend and further refining models by including terms such as tropical upper-tropospheric heating. This also indicates that the current decline in sea ice cover may not result in a more negative NAO. In contrast, the GloSea5 predictions of the winter NAO show no such bias and a closer correspondence to the observed NAO from 2014 to 2016 (forecasts were not issued in 2013).

While models differ in the precise predictors selected, there is a broad similarity among the predictors, indicating greater confidence that the association with predictors is genuine, rather than fitting to noise. N3.4 is used in all models and a sea ice term appears in all models apart from N56. The relationship with the North Atlantic June tripole is found in N80 only, while an extratropical Atlantic influence is also present in N93 (the AMO), although a recent study suggests that the main AMO influence on the North Atlantic atmospheric circulation comes from tropical SSTs (Davini et al. 2015). The only suggestion of solar variability influence is in the longer series N56, and the west Indian Ocean influence is indicated in all three statistical models. The influence of these predictors is confirmed in modeling studies (e.g., Li et al. 2010; Maidens et al. 2013; Andrews et al. 2015) and therefore suggests that genuine skill is present in the statistical forecasts. However, as the statistical models use a limited range of predictors only, there are likely to be periods when they are less successful than dynamical forecasts such as those from GloSea5, when other factors may be more dominant.

Also of interest are the predictors that are not selected by the models. Despite the available evidence (e.g., Ebdon 1975), relationships between the QBO and winter NAO were not found to be strong enough to warrant inclusion in the models. While studies with dynamical models suggest the need for a fully resolved stratosphere (e.g., Marshall and Scaife 2010; Scaife et al. 2016), the stratospheric influence via the drivers selected in these statistical models is limited, probably just to N3.4 (Bell et al. 2009). Similarly, no role for Eurasian snow cover is identified, despite evidence presented in other research (Cohen and Jones 2011; Riddle et al. 2013). Although Cohen and Jones (2011) found their snow advance index (SAI) demonstrated better correlation with winter NAO than did snow cover extent, the reason for this has not been established and their 2014 forecasts were poor. Here, snow cover shows covariance with sea ice in the Barents–Kara Sea and is thus not selected.

The detrended sea ice forecast model compares favorably to GloSea5 when used for out-of-sample forecasts (correlation of 0.56 compared with 0.61 for GloSea5). The ability of a statistical forecast to correctly predict the winter NAO means that such forecasts can act as benchmarks for dynamical forecast models such as GloSea5. For example, simple statistical models may shed light on the reasons why a dynamical model issues a poor forecast in particular years by identifying a particular factor. The poor hindcast of winter 2005 in GloSea5 has an as yet undetermined cause but the skill in the statistical models may come from tropical rainfall. While the signal from tropical rainfall was evident in GloSea5 and suggestive of a hindcast similar to statistical models, other as yet unidentified processes within the model prevented this. The statistical approach used lends support to the argument that the winter NAO has a significant predictable component and shows that skillful and reliable statistical forecasts are possible. In the future, these simple forecasts can be extended to incorporate other predictors and nonlinear relationships through the use of more advanced methods.

Acknowledgments. AAS was supported by the Joint DECC/Defra Met Office Hadley Centre Programme (GA1101). RJH, JMJ, EH, and RE acknowledge funding support from the University of Sheffield Project Sunshine. RE is grateful to NSF, Hungary (OTKA; Reference K83133). We thank the three anonymous reviewers whose comments helped to significantly enhance the text.

REFERENCES

Adler, R. F., and Coauthors, 2003: The version-2 Global Precipitation Climatology Project (GPCP) monthly precipitation analysis (1979–present). *J. Hydrometeor.*, **4**, 1147–1167, doi:10.1175/1525-7541(2003)004<1147:TVGPCP>2.0.CO;2.

Anderson, T. W., and D. A. Darling, 1952: Asymptotic theory of certain ''goodness of fit'' criteria based on stochastic processes. *Ann. Math. Stat.*, **23**, 193–212, doi:10.1214/aoms/1177729437.

Andrews, M. B., J. R. Knight, and L. J. Gray, 2015: A simulated lagged response of the North Atlantic Oscillation to the solar cycle over the period 1960–2009. *Environ. Res. Lett.*, **10**, 054022, doi:10.1088/1748-9326/10/5/054022.

Anstey, J. A., and T. G. Shepherd, 2014: High-latitude influence of the quasi-biennial oscillation. *Quart. J. Roy. Meteor. Soc.*, **140**, 1–21, doi:10.1002/qj.2132.

Arribas, A., and Coauthors, 2011: The GloSea4 ensemble prediction system for seasonal forecasting. *Mon. Wea. Rev.*, **139**, 1891–1910, doi:10.1175/2010MWR3615.1.

Bader, J., and M. Latif, 2003: The impact of decadal-scale Indian Ocean sea surface temperature anomalies on Sahelian rainfall and the North Atlantic Oscillation. *Geophys. Res. Lett.*, **30**, 2169, doi:10.1029/2003GL018426.

Baldwin, M. P., and T. J. Dunkerton, 2001: Stratospheric harbingers of anomalous weather regimes. *Science*, **294**, 581–584, doi:10.1126/science.1063315.

Barnes, E. A., and L. M. Polvani, 2015: CMIP5 projections of Arctic amplification, of the North American/North Atlantic circulation, and of their relationship. *J. Climate*, **28**, 5254–5271, doi:10.1175/JCLI-D-14-00589.1.

——, and J. A. Screen, 2015: The impact of Arctic warming on the midlatitude jet-stream: Can it? Has it? Will it? *Wiley Interdiscip. Rev: Climatic Change*, **6**, 277–286, doi:10.1002/wcc.337.

Bell, C. J., L. J. Gray, A. J. Charlton-Perez, and M. M. Joshi, 2009: Stratospheric communication of El Niño teleconnections to European winter. *J. Climate*, **22**, 4083–4096, doi:10.1175/2009JCLI2717.1.

Bröcker, J., and L. A. Smith, 2007: Increasing the reliability of reliability diagrams. *Wea. Forecasting*, **22**, 651–661, doi:10.1175/WAF993.1.

Butler, A. H., D. W. J. Thompson, and R. Heikes, 2010: The steady-state atmospheric circulation response to climate change–like thermal forcings in a simple general circulation model. *J. Climate*, **23**, 3474–3496, doi:10.1175/2010JCLI3228.1.

Cavalieri, D., C. Parkinson, P. Gloerson, and H. J. Zwally, 1996: Sea ice concentrations from Nimbus-7 SMMR and DMSP SSM/I-SSMIS passive microwave data, version 1. NASA DAAC, National Snow and Ice Data Center, Subset used: North/monthly/1979–2015, accessed 13 November 2015, doi:10.5067/8GQ8LZQVL0VL.

Choulakian, V., R. A. Lockhart, and M. A. Stephens, 1994: Cramér–von Mises statistics for discrete distributions. *Can. J. Stat.*, **22**, 125–137, doi:10.2307/3315828.

Cohen, J., and J. Jones, 2011: A new index for more accurate winter predictions. *Geophys. Res. Lett.*, **38**, L21701, doi:10.1029/2011GL049626.

Czaja, A., and J. Marshall, 2001: Observations of atmosphere–ocean coupling in the North Atlantic. *Quart. J. Roy. Meteor. Soc.*, **127**, 1893–1916, doi:10.1002/qj.49712757603.

Davini, P., J. von Hardenberg, and S. Corti, 2015: Tropical origin for the impacts of the Atlantic Multidecadal Variability on the Euro-Atlantic climate. *Environ. Res. Lett.*, **10**, 094010, doi:10.1088/1748-9326/10/9/094010.

Deser, C., M. A. Alexander, and M. S. Timlin, 2003: Understanding the persistence of sea surface temperature anomalies in midlatitudes. *J. Climate*, **16**, 57–72, doi:10.1175/1520-0442(2003)016<0057:UTPOSS>2.0.CO;2.

Eade, R., D. Smith, A. A. Scaife, E. Wallace, N. Dunstone, L. Hermanson, and N. Robinson, 2014: Do seasonal-to-decadal climate predictions underestimate the predictability of the real world? *Geophys. Res. Lett.*, **41**, 5620–5628, doi:10.1002/2014GL061146.

Ebdon, R. A., 1975: The quasi-biennial oscillation and its association with tropospheric circulation patterns. *Meteor. Mag.*, **104**, 282–297.

Elmore, K., 2005: Alternatives to the chi-square test for evaluating rank histograms from ensemble forecasts. *Wea. Forecasting*, **20**, 789–795, doi:10.1175/WAF884.1.

Enfield, D. B., A. M. Mestas-Nuñez, and P. J. Trimble, 2001: The Atlantic Multidecadal Oscillation and its relationship to rainfall and river flows in the continental U.S. *Geophys. Res. Lett.*, **28**, 2077–2080, doi:10.1029/2000GL012745.

Feldstein, S., 2000: The timescale, power spectra, and climate noise properties of teleconnection patterns. *J. Climate*, **13**, 4430–4440, doi:10.1175/1520-0442(2000)013<4430:TTPSAC>2.0.CO;2.

——, 2002: The recent trend and variance increase of the annular mode. *J. Climate*, **15**, 88–94, doi:10.1175/1520-0442(2002)015<0088:TRTAVI>2.0.CO;2.

Folland, C. K., A. A. Scaife, J. Lindesay, and D. B. Stephenson, 2012: How potentially predictable is northern European winter climate a season ahead? *Int. J. Climatol.*, **32**, 801–818, doi:10.1002/joc.2314.

Gray, L. J., and Coauthors, 2013: A lagged response to the 11 year solar cycle in observed winter Atlantic/European weather patterns. *J. Geophys. Res. Atmos.*, **118**, 13 405–13 420, doi:10.1002/2013JD020062.

Hall, R., R. Erdélyi, E. Hanna, J. M. Jones, and A. A. Scaife, 2015: Drivers of North Atlantic polar front jet stream variability. *Int. J. Climatol.*, **35**, 1697–1720, doi:10.1002/joc.4121.

Hamilton, K., 1984: Mean wind evolution through the quasi-biennial cycle in the tropical lower stratosphere. *J. Atmos. Sci.*, **41**, 2113–2125, doi:10.1175/1520-0469(1984)041<2113:MWETTQ>2.0.CO;2.

Hanna, E., T. E. Cropper, P. D. Jones, A. A. Scaife, and R. Allan, 2015: Recent seasonal asymmetric changes in the NAO (a marked summer decline and increased winter variability) and associated changes in the AO and Greenland blocking index. *Int. J. Climatol.*, **35**, 2540–2554, doi:10.1002/joc.4157.

Hermanson, L., R. Eade, N. H. Robinson, N. J. Dunstone, M. B. Andrews, J. R. Knight, A. A. Scaife, and D. M. Smith, 2014: Forecast cooling of the Atlantic subpolar gyre and associated impacts. *Geophys. Res. Lett.*, **41**, 5167–5174, doi:10.1002/2014GL060420.

Hoerling, M. P., J. W. Hurrell, T. Xu, G. T. Bates, and A. S. Phillips, 2004: Twentieth century North Atlantic climate change. Part II: Understanding the effect of Indian Ocean warming. *Climate Dyn.*, **23**, 391–405, doi:10.1007/s00382-004-0433-x.

Holland, M. M., D. A. Bailey, and S. Vavrus, 2011: Inherent predictability in the rapidly changing Arctic environment of the Community Climate System Model, version 3. *Climate Dyn.*, **36**, 1239–1253, doi:10.1007/s00382-010-0792-4.

Holton, J. R., and H.-C. Tan, 1980: The influence of the equatorial quasi-biennial oscillation on the global circulation at 50 mb. *J. Atmos. Sci.*, **37**, 2200–2208, doi:10.1175/1520-0469(1980)037<2200:TIOTEQ>2.0.CO;2.

Hoskins, B. J., and D. J. Karoly, 1981: The steady linear response of a spherical atmosphere to thermal and orographic forcing. *J. Atmos. Sci.*, **38**, 1179–1196, doi:10.1175/1520-0469(1981)038<1179:TSLROA>2.0.CO;2.

Hurrell, J. W., 1995: Decadal trends in the North Atlantic Oscillation: Regional temperatures and precipitation. *Science*, **269**, 676–679, doi:10.1126/science.269.5224.676.

——, and C. Deser, 2009: North Atlantic climate variability: The role of the North Atlantic Oscillation. *J. Mar. Syst.*, **78**, 28–41, doi:10.1016/j.jmarsys.2008.11.026.

——, Y. Kushnir, G. Ottersen, and M. Visbeck, 2003: *The North Atlantic Oscillation: Climate Significance and Environmental Impact.* Amer. Geophys. Union, 279 pp.

Ineson, S., A. A. Scaife, J. R. Knight, J. C. Manners, N. J. Dunstone, L. J. Gray, and J. D. Haigh, 2011: Solar forcing of winter climate variability in the Northern Hemisphere. *Nat. Geosci.*, **4**, 753–757, doi:10.1038/ngeo1282.

James, I. N., and P. M. James, 1989: Ultra-low-frequency variability in a simple atmospheric circulation mode. *Nature*, **342**, 53–55, doi:10.1038/342053a0.

Jung, T., F. Vitart, L. Ferranti, and J.-J. Morcrette, 2011: Origin and predictability of the extreme negative NAO winter of 2009/10. *Geophys. Res. Lett.*, **38**, L07701, doi:10.1029/2011GL046786.

Kaplan, A., M. Cane, Y. Kushnir, A. Clement, M. Blumenthal, and B. Rajagopalan, 1998: Analyses of global sea surface temperature 1856–1991. *J. Geophys. Res.*, **103**, 18 567–18 589, doi:10.1029/97JC01736.

Kay, J. E., M. M. Holland, and A. Jahn, 2011: Inter-annual to multi-decadal Arctic sea ice extent trends in a warming world. *Geophys. Res. Lett.*, **38**, L15708, doi:10.1029/2011GL048008.

Keeley, S. P. E., R. T. Sutton, and L. C. Shaffrey, 2009: Does the North Atlantic Oscillation show unusual persistence on intraseasonal timescales? *Geophys. Res. Lett.*, **36**, L22706, doi:10.1029/2009GL040367.

Kim, H.-M., P. Webster, and J. Curry, 2012: Seasonal prediction skill of ECMWF system 4 and NCEP CFSv2 retrospective forecast for the Northern Hemisphere winter. *Climate Dyn.*, **39**, 2957–2973, doi:10.1007/s00382-012-1364-6.

Li, S., J. Perlwitz, M. P. Hoerling, and X. Chen, 2010: Opposite annular responses of the Northern and Southern Hemispheres to Indian Ocean warming. *J. Climate*, **23**, 3720–3738, doi:10.1175/2010JCLI3410.1.

MacLachlan, C., and Coauthors, 2014: Global Seasonal Forecast System version 5 (GloSea5): A high-resolution seasonal forecast system. *Quart. J. Roy. Meteor. Soc.*, **141**, 1072–1084, doi:10.1002/qj.2396.

Maidens, A., A. Arribas, A. A. Scaife, C. MacLachlan, D. Peterson, and J. Knight, 2013: The influence of surface forcings on prediction of the North Atlantic Oscillation regime of winter 2010/11. *Mon. Wea. Rev.*, **141**, 3801–3813, doi:10.1175/MWR-D-13-00033.1.

Marshall, A. G., and A. A. Scaife, 2010: Improved predictability of stratospheric sudden warming events in an atmospheric general circulation model with enhanced stratospheric resolution. *J. Geophys. Res.*, **115**, D16114, doi:10.1029/2009JD012643.

Mason, S. J., and N. E. Graham, 2002: Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation. *Quart. J. Roy. Meteor. Soc.*, **128**, 2145–2166, doi:10.1256/003590002320603584.

Naujokat, B., 1986: An update of the observed quasi-biennial oscillation of the stratospheric winds over the tropics. *J. Atmos. Sci.*, **43**, 1873–1877, doi:10.1175/1520-0469(1986)043<1873:AUOTOQ>2.0.CO;2.

Park, H.-S., S. Lee, S.-W. Son, S. B. Feldstein, and Y. Kosaka, 2015: The impact of poleward moisture and sensible heat flux on Arctic winter sea ice variability. *J. Climate*, **28**, 5030–5040, doi:10.1175/JCLI-D-15-0074.1.

Rayner, N. A., D. E. Parker, E. B. Horton, C. K. Folland, L. V. Alexander, D. P. Rowell, E. C. Kent, and A. Kaplan, 2003: Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *J. Geophys. Res.*, **108**, 4407, doi:10.1029/2002JD002670.

Riddle, E. E., A. H. Butler, J. C. Furtado, J. L. Cohen, and A. Kumar, 2013: CFSv2 ensemble prediction of the wintertime Arctic Oscillation. *Climate Dyn.*, **41**, 1099–1116, doi:10.1007/s00382-013-1850-5.

Robinson, D. A., T. W. Estilow, and NOAA CDR Program, 2012: NOAA Climate Data Record (CDR) of Northern Hemisphere (NH) snowcover extent (SCE), v01r01. NOAA/National Climatic Data Center, Subset used: 1979–2015, accessed 13 February 2015, doi:10.7289/V5N014G9.

Robock, A., and J. Mao, 1995: The volcanic signal in surface temperature observations. *J. Climate*, **8**, 1086–1103, doi:10.1175/1520-0442(1995)008<1086:TVSIST>2.0.CO;2.

Robson, J., R. Sutton, K. Lohmann, D. Smith, and M. D. Palmer, 2012: Causes of the rapid warming of the North Atlantic Ocean in the mid-1990s. *J. Climate*, **25**, 4116–4134, doi:10.1175/JCLI-D-11-00443.1.

Rodwell, M. J., and C. K. Folland, 2002: Atlantic air–sea interaction and seasonal predictability. *Quart. J. Roy. Meteor. Soc.*, **128**, 1413–1443, doi:10.1002/qj.200212858302.

——, D. P. Rowell, and C. K. Folland, 1999: Oceanic forcing of the wintertime North Atlantic Oscillation and European climate. *Nature*, **398**, 320–323, doi:10.1038/18648.

Sato, K., J. Inoue, and M. Watanabe, 2014: Influence of the Gulf Stream on the Barents Sea ice retreat and Eurasian coldness during early winter. *Environ. Res. Lett.*, **9**, 084009, doi:10.1088/1748-9326/9/8/084009.

Scaife, A. A., and Coauthors, 2009: The CLIVAR C20C project: Selected twentieth century climate events. *Climate Dyn.*, **33**, 603–614, doi:10.1007/s00382-008-0451-1.

——, S. Ineson, J. R. Knight, L. Gray, K. Kodera, and D. M. Smith, 2013: A mechanism for lagged North Atlantic climate response to solar variability. *Geophys. Res. Lett.*, **40**, 434–439, doi:10.1002/grl.50099.

——, and Coauthors, 2014: Skillful long-range prediction of European and North American winters. *Geophys. Res. Lett.*, **41**, 2514–2519, doi:10.1002/2014GL059637.

——, and Coauthors, 2016: Seasonal winter forecasts and the stratosphere. *Atmos. Sci. Lett.*, **17**, 51–56, doi:10.1002/asl.598.

——, and Coauthors, 2017: Predictability of European winter 2015/16. *Atmos. Sci. Lett.*, **18**, 38–44, doi:10.1002/asl.721.

Stenchikov, G., L. Hamilton, R. J. Stouffer, R. Robock, V. Ramaswamy, B. Santer, and H.-F. Graf, 2006: Arctic Oscillation response to volcanic eruptions in the IPCC AR4 climate models. *J. Geophys. Res.*, **111**, D07107, doi:10.1029/2005JD006286.

Stroeve, J. C., V. Kattsov, A. Barrett, M. Serreze, T. Pavlova, M. Holland, and W. N. Meier, 2012: Trends in Arctic sea ice extent from CMIP5, CMIP3 and observations. *Geophys. Res. Lett.*, **39**, L16502, doi:10.1029/2012GL052676.

——, A. D. Crawford, and S. Stammerjohn, 2016: Using timing of ice retreat to predict timing of fall freeze-up in the Arctic. *Geophys. Res. Lett.*, **43**, 6332–6340, doi:10.1002/2016GL069314.

Strong, C., and G. Magnusdottir, 2011: Dependence of NAO variability on coupling with sea ice. *Climate Dyn.*, **36**, 1681–1689, doi:10.1007/s00382-010-0752-z.

Toniazzo, T., and A. A. Scaife, 2006: The influence of ENSO on winter North Atlantic climate. *Geophys. Res. Lett.*, **33**, L24704, doi:10.1029/2006GL027881.

Vallis, G. K., and E. P. Gerber, 2008: Local and hemispheric dynamics of the North Atlantic Oscillation, annular patterns and the zonal index. *Dyn. Atmos. Oceans*, **44**, 184–212, doi:10.1016/j.dynatmoce.2007.04.003.

Watson, G. S., 1961: Goodness of fit tests on a circle. *Biometrika*, **48**, 109–114, doi:10.1093/biomet/48.1-2.109.

Wilks, D. S., 2006: Comparison of ensemble-MOS methods in the Lorenz '96 setting. *Meteor. Appl.*, **13**, 243–256, doi:10.1017/S1350482706002192.

——, 2011: *Statistical Methods for the Atmospheric Sciences*. 3rd ed. Elsevier, 676 pp.

Woollings, T., and M. Blackburn, 2012: The North Atlantic jet stream under climate change and its relation to the NAO and EA patterns. *J. Climate*, **25**, 886–902, doi:10.1175/JCLI-D-11-00087.1.